Medical Text Mining: A DLI-2 Status Report

Hsinchun Chen

McClelland Professor,

Director,

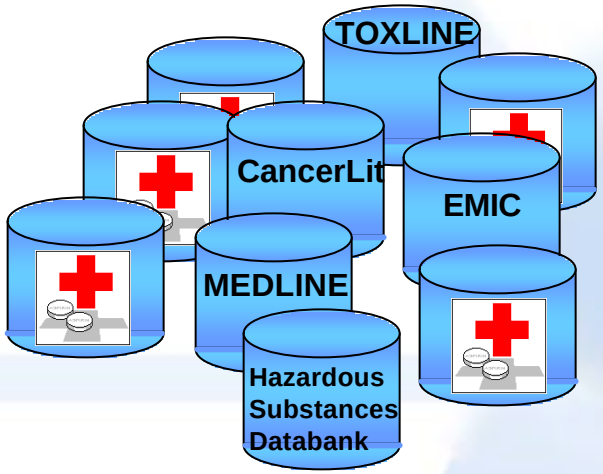
Artificial Intelligence Lab and Hoffman eCommerce Lab

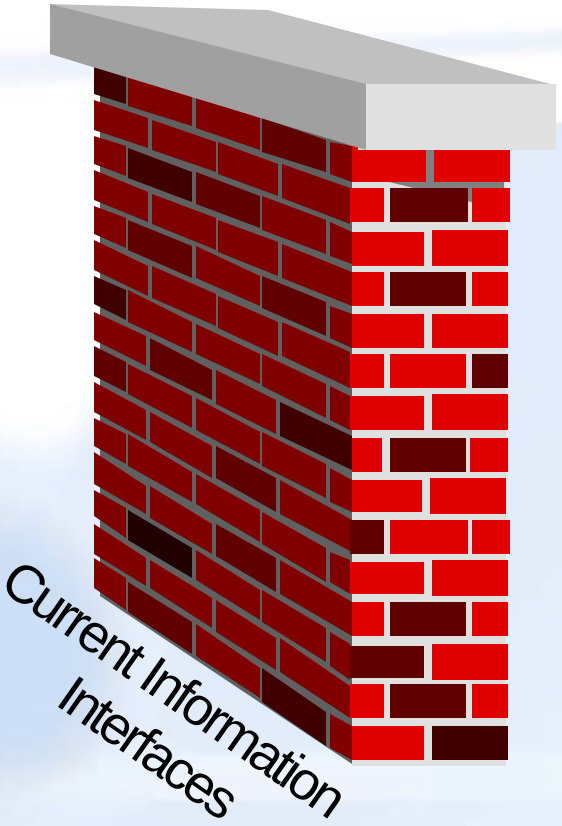
The University of Arizona

Acknowledgement:
NSF DLI2, NIH/NLM

The Medical Information Gap

Heterogeneous Medical Literature Databases and the Internet





Medical Professionals & Users

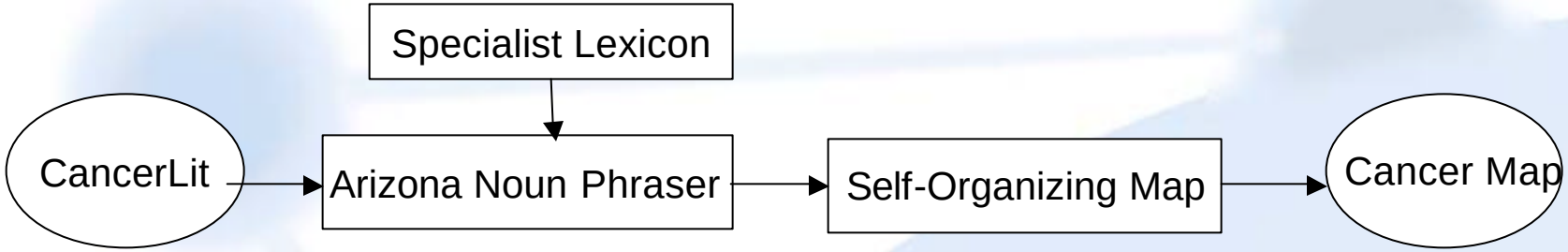
Research Questions in Medical Text Mining

- How can linguistic parsing and statistical analysis techniques help extract medical terminology and the relationships between terms?
- How can medical and general ontologies help improve extraction of medical terminology?
- How can linguistic parsing, statistical analysis, and ontologies be incorporated in customizable retrieval interfaces?

Medical Text Mining Research Update

- Cancer Map: Knowledge Map for Cancer Researchers
- Cancer Spider: Client-based Meta Cancer Search Agent
- Chinese MED Map: Multilingual Research in Medical Informatics

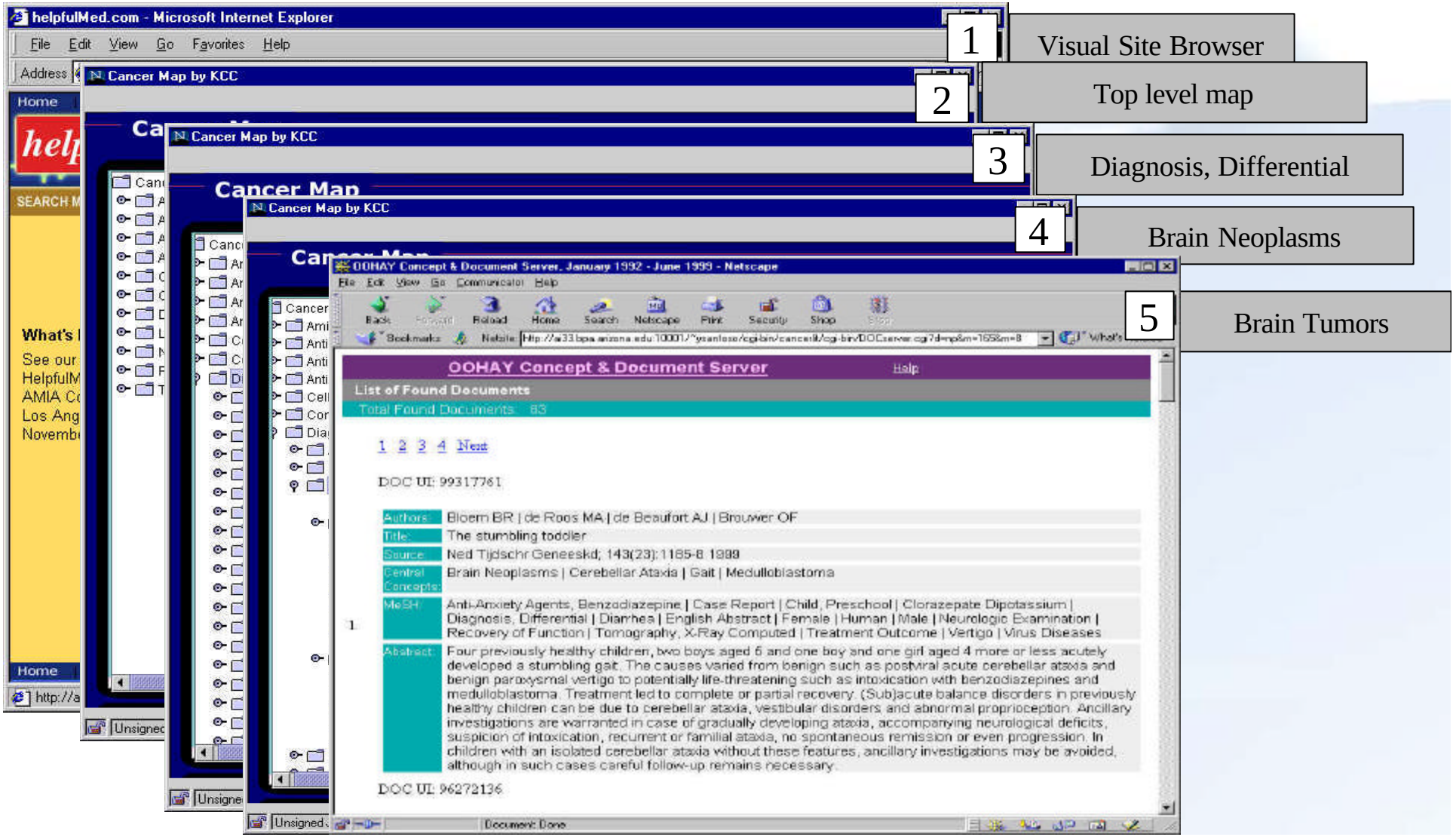
Cancer Map System Architecture



- Linguistic analysis: Arizona Noun Phraser and UMLS Specialist Lexicon
- Kohonen Self-Organizing Map (SOM): Topic and document categorization

- Multi-layered graphical display of important cancer concepts supports *browsing* of cancer literature (1M+ CancerLit documents)
- Presents 21,000 cancer topics on 1180 maps organized in 5 layers

Browsing Cancer Map



1. Visual Site Browser
2. Top level map
3. Diagnosis, Differential
4. Brain Neoplasms
5. Brain Tumors

Cancer Map User Study

- Compare topic hierarchies on Cancer Map vs. topic hierarchies on MeSH cancer subtree
- 30 cancer researchers from Arizona Cancer Center as subjects

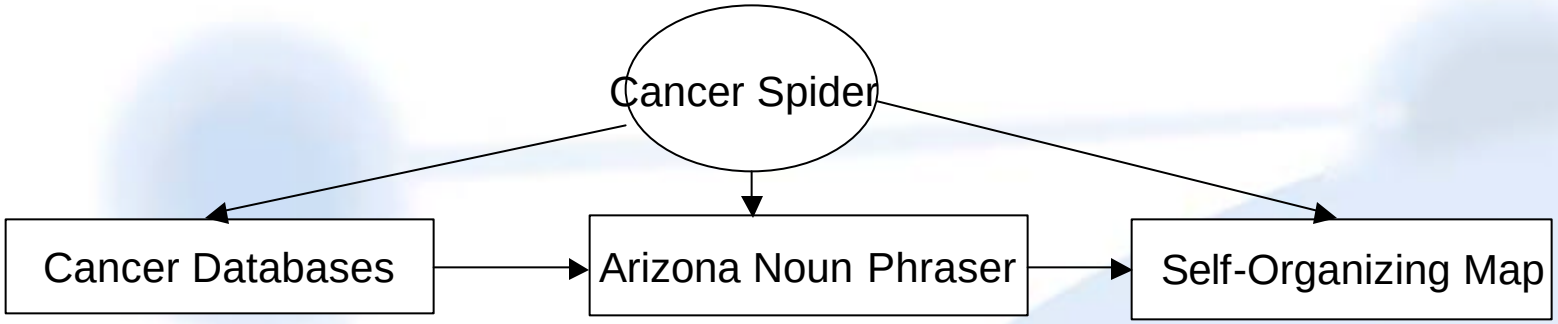
- Future work in interface comparison: size, proximity, and layers (Topic Island)

Cancer Map User Study Result

C: Cancer Map M: MeSH	First level	Second level (overlap)	Second level (non-overlap)	Third level
Recall Comparison	C: 0.557	C: 0.765	C: 0.859	C: 0.839
	M: 0.466	M: 0.113	M: 0.466	M: 0.459
	P = 0.049	**P = 0.00**	**P = 0.000**	**P = 0.003**
Precision Comparison	C: 0.926	C: 0.826	C: 0.829	C: 0.863
	M: 0.956	M: 0.608	M: 0.904	M: 0.917
	P = 0.591	**P = 0.104**	**P = 0.459**	**P = 0.808**

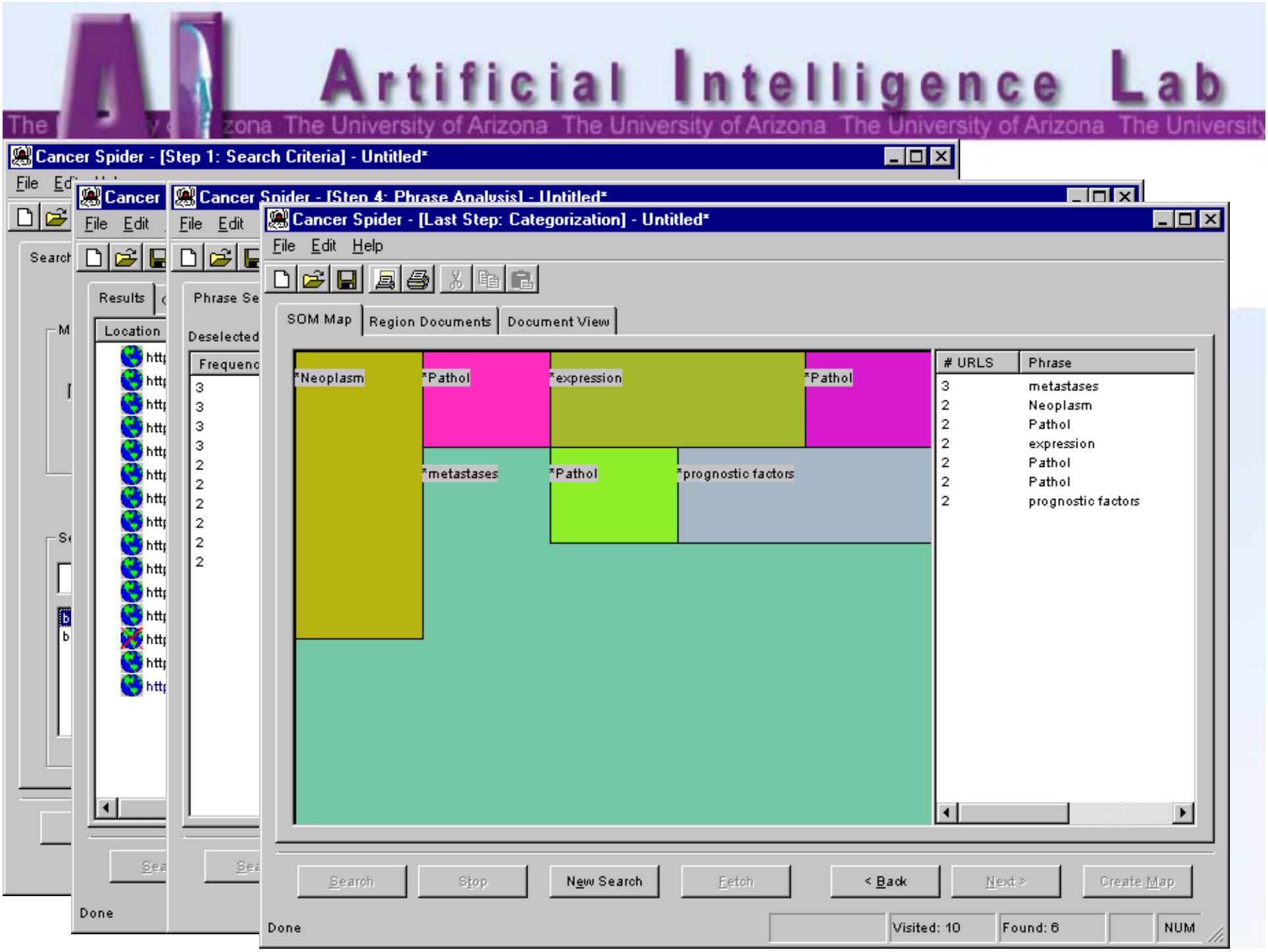
- Cancer Map was comparable to MeSH cancer subtree in *perceived topic precision* at each level and was significantly better than MeSH in *perceived topic recall* at all levels.
- Cancer Map and MeSH are complementary in topic suggestion.

Cancer Spider System Architecture



- Cancer databases: CancerLit, PDQ, Medline
- Linguistic analysis: Arizona Noun Phraser and UMLS Specialist Lexicon
- Kohonen Self-Organizing Map (SOM): Topic and document categorization

- Meta searching: Search multiple cancer databases
- Iterative, dynamic, personalized medical topic/theme summarization


Artificial Intelligence Lab
The University of Arizona
Cancer Spider - [Step 1: Search Criteria] - Untitled*
Cancer Spider - [Step 4: Phrase Analysis] - Untitled*
Cancer Spider - [Last Step: Categorization] - Untitled*
File Edit Help
SOM Map
Region Documents
Document View
*Neoplasm
*Pathol
*expression
*Pathol
*metastases
*Pathol
*prognostic factors
URLS
Phrase
3 metastases
2 Neoplasm
2 Pathol
2 expression
2 Pathol
2 Pathol
2 prognostic factors
Search
Stop
New Search
Fetch
< Back
Next >
Create Map
Done
Visited: 10
Found: 6
NUM

Cancer Spider User Study

- Compare Cancer Spider vs. NLM Gateway (NLM's one-stop MED portal with MeSH thesaurus)

  http://gateway.nlm.nih.gov/gw/Cmd?GMBasicSearch
- 30 cancer researchers from Arizona Cancer Center as subjects; 2 expert evaluators
- Performance measures: document recall, precision, F, time spent, # of documents

Cancer Spider User Study Result

	Sample size	CancerSpider		NLM Gateway		P-Value
		Mean	Variance	Mean	Variance	
Precision	30	0.803	0.117	0.826	0.122	0.7572
Recall	30	0.533	0.112	0.539	0.121	0.9523
F-measure	30	0.612	0.105	0.622	0.110	0.9056

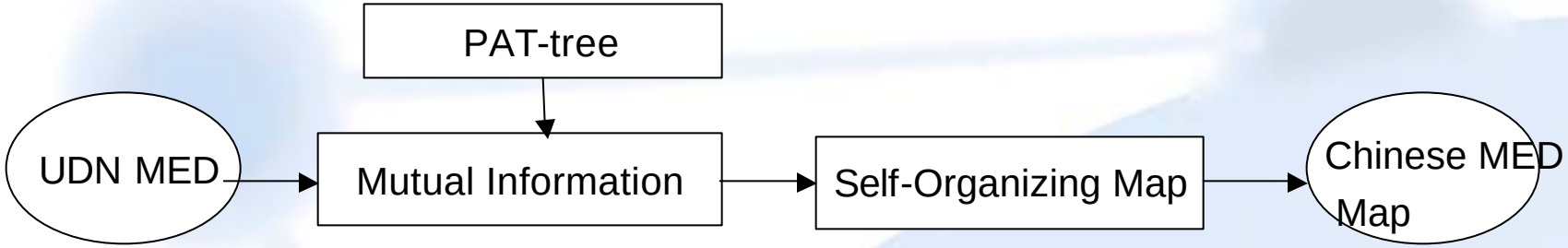
•Cancer Spider and NLM Gateway users achieved similar performances in precision, recall, and F measure.

Cancer Spider User Study Result

	Sample size	CancerSpider		NLM Gateway		P-Value
		Mean	Variance	Mean	Variance	
Time (in minutes)	30	**10.22**	15.64	14.00	23.18	0.0003
No. of documents browsed	30	**4.533**	3.586	6.233	12.461	0.0067

- Cancer Spider users spent less time and browsed fewer documents**.**
- The two tools are complementary in functionalities and documents suggested.

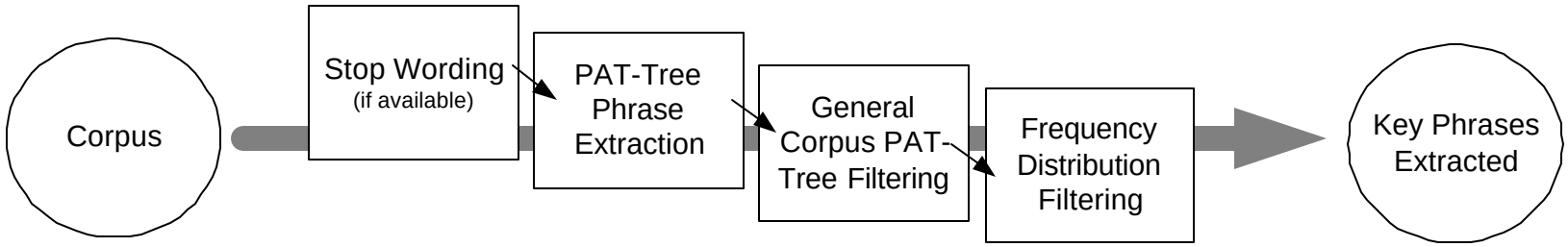
Chinese MED Map System Architecture



- Statistic-based indexing: Mutual Information and PAT-tree
- Kohonen Self-Organizing Map (SOM): Topic and document categorization

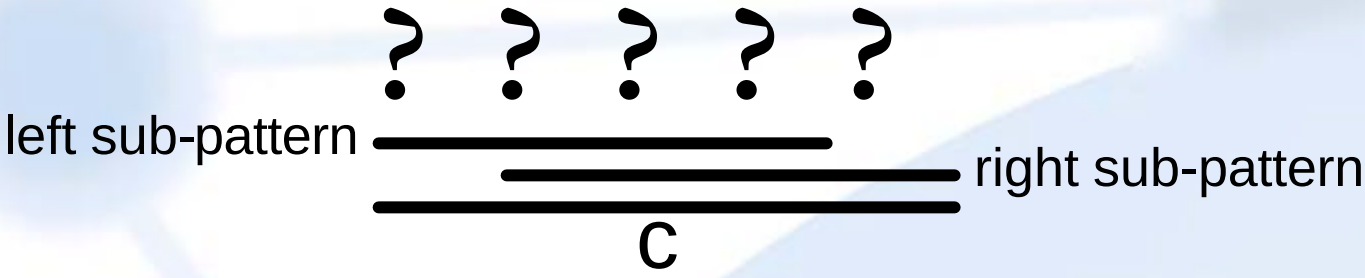
- Multi-layered graphical display of important health-related topics supports *browsing* of health-related news articles (70K+ United Daily News, UDN MED)
- User study underway

Chinese Medical Indexing Architecture



- Stop wording breaks long sentence into smaller chunks to reduce noise
- "Updateable" mutual information technique improves precision of extraction
- General PAT-tree filter general phrases
- Frequency distribution filtering distills less useful terms

Mutual Information (MI) estimator



$$MI_c \;?\; \frac{f_c}{f_{left} \;?\; f_{right} \;?\; f_c}$$

$$MI_c^{?} \;?\; \log_2 \frac{\Pr(c)}{\Pr(left)\Pr(right)}$$

Note: $|A \;?\; B| \;?\; |A| \;?\; |B| \;?\; |A \;?\; B|$

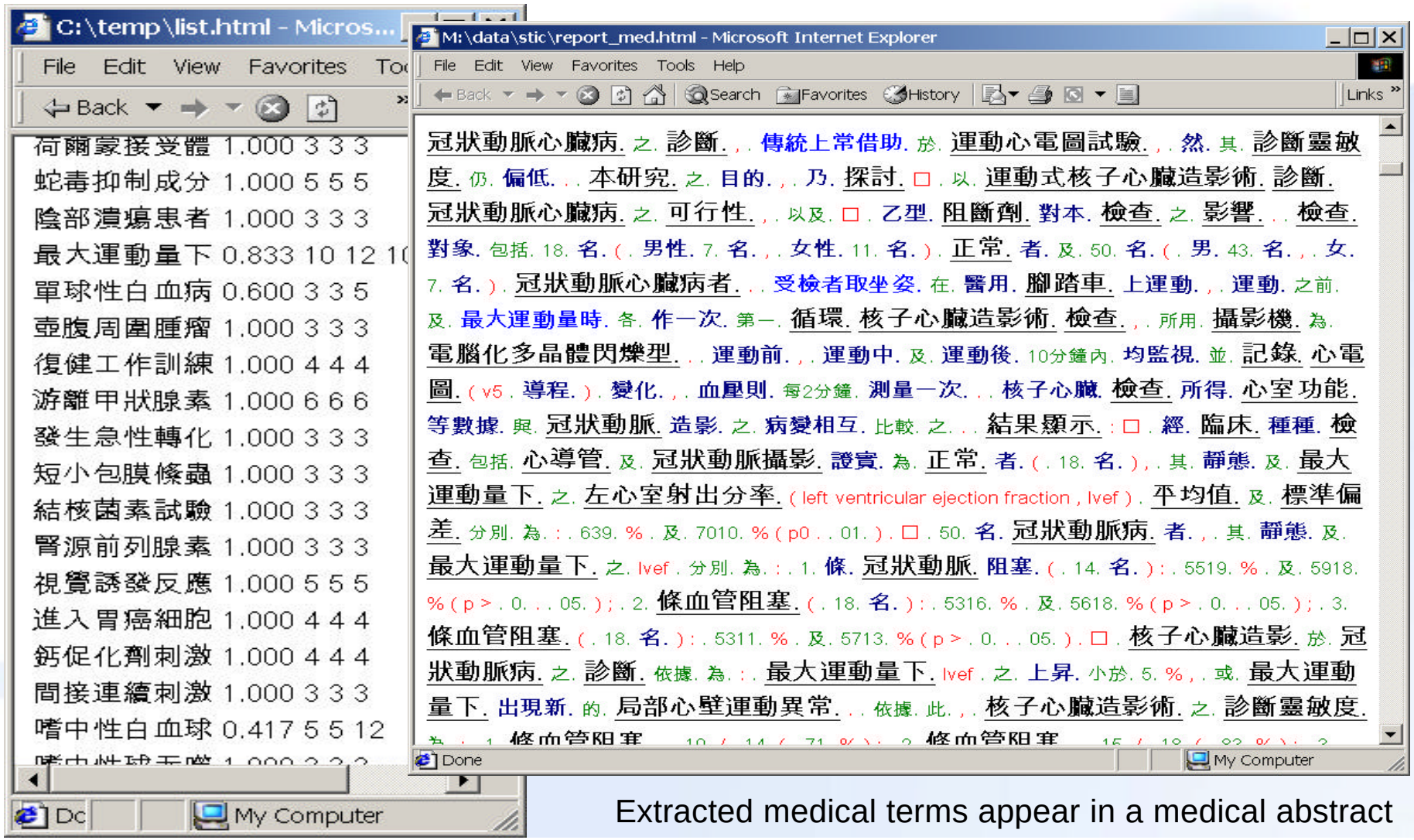
Independent event: $\Pr(A \;?\; B) \;?\; \Pr(A)\Pr(B)$

MI_c = 0 no correlation

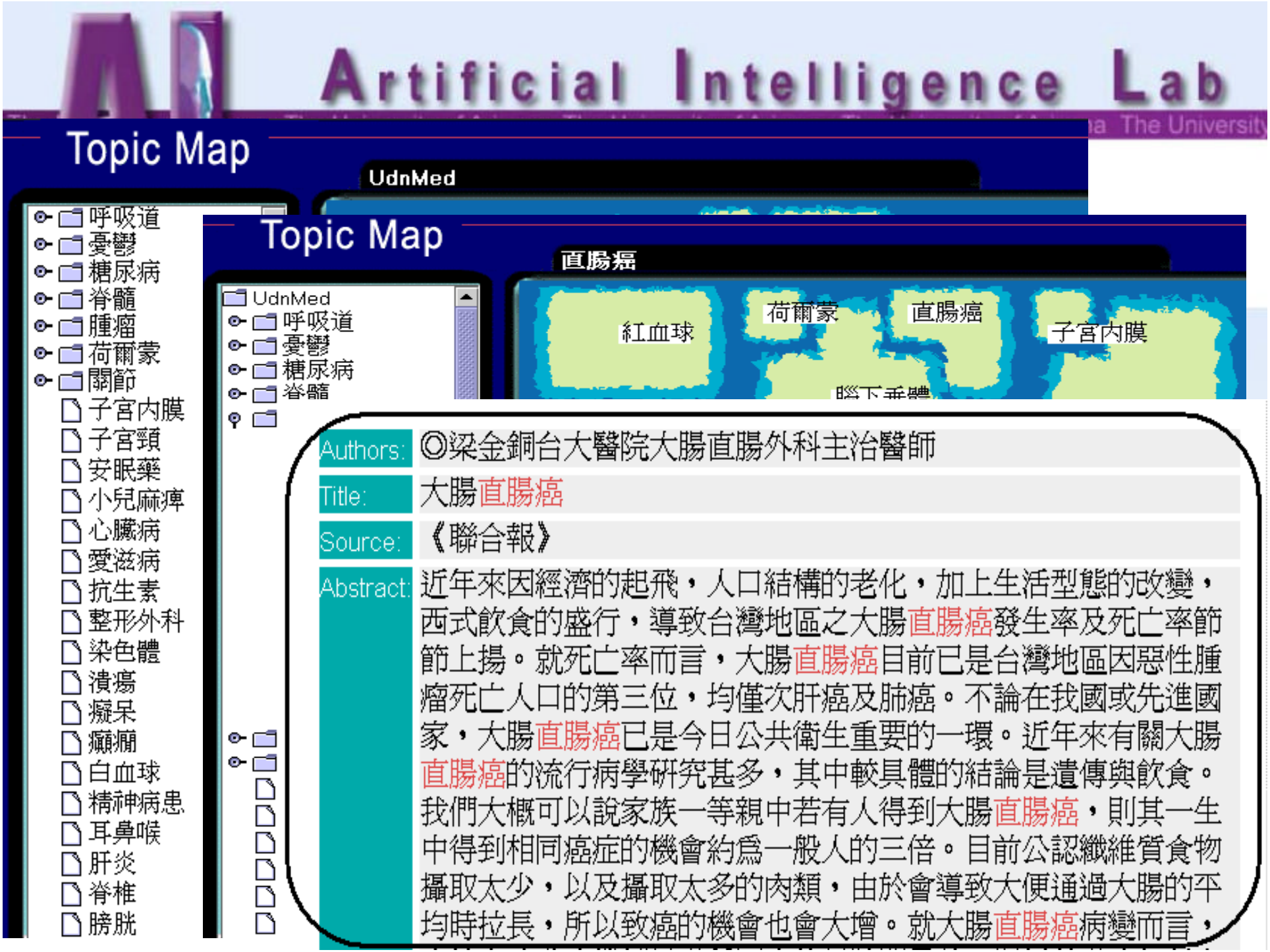
MI_c = 1 perfect correlation

List of extracted phrases

Sample Medical Abstract



Extracted medical terms appear in a medical abstract


Artificial Intelligence Lab
The University
Topic Map
UdnMed
呼吸道
憂鬱
糖尿病
脊髓
腫瘤
荷爾蒙
關節
子宮內膜
子宮頸
安眠藥
小兒麻痺
心臟病
愛滋病
抗生素
整形外科
染色體
潰瘍
癡呆
癲癇
白血球
精神病患
耳鼻喉
肝炎
脊椎
膀胱
Topic Map
UdnMed
呼吸道
憂鬱
糖尿病
脊髓
直腸癌
紅血球
荷爾蒙
直腸癌
子宮內膜
Authors:
◎梁金銅台大醫院大腸直腸外科主治醫師
Title:
大腸直腸癌
Source:
《聯合報》
Abstract:
近年來因經濟的起飛，人口結構的老化，加上生活型態的改變，西式飲食的盛行，導致台灣地區之大腸直腸癌發生率及死亡率節節上揚。就死亡率而言，大腸直腸癌目前已是台灣地區因惡性腫瘤死亡人口的第三位，均僅次肝癌及肺癌。不論在我國或先進國家，大腸直腸癌已是今日公共衛生重要的一環。近年來有關大腸直腸癌的流行病學研究甚多，其中較具體的結論是遺傳與飲食。我們大概可以說家族一等親中若有人得到大腸直腸癌，則其一生中得到相同癌症的機會約爲一般人的三倍。目前公認纖維質食物攝取太少，以及攝取太多的肉類，由於會導致大便通過大腸的平均時拉長，所以致癌的機會也會大增。就大腸直腸癌病變而言，

Future Research Directions

- Text mining + MED Ontologies
- Information visualization for medical informatics
- Multilingual medical informatics: text mining + Multilingual UMLS
- GeneScene: Gene pathway analysis and visualization

Research Announcement:

- ICADL2001, International Conference of Asian DL, Bangalore, India, 12/10-12/12/ 2001 (paper due: 7/15/2001)
- CFP, Special Topics Issue of DSS: “Web Retrieval and Mining”, 7/13/2001
- CFP, Special Topics Issue of DSS: “Digital Government: Technologies and Practices”, 8/14/2001
- CFP, Special Topics Issue of JASIST: “Web Retrieval and Mining”, 9/14/2001

For Project Information:

http://ai.bpa.arizona.edu

Hchen@bpa.arizona.edu

For Medical Demos:

http://www.HelpfulMED.com